QDETRv: Query-Guided DETR for One-Shot Object Localization in Videos

Yogesh Kumar[1], **Saswat Mallick**[1], **Anand Mishra**[1], **Sowmya Rasipuram**[2],
Anutosh Maitra[2], **Roshni Ramnani**[2]
[1]Indian Institute of Technology Jodhpur, India
[2]Accenture Labs
kumar.204@iitj.ac.in, saswatsubhajyotimallick@gmail.com, mishra@iitj.ac.in,
{sowmya.rasipuram, anutosh.maitra, roshni.r.ramnani}@accenture.com

Abstract

In this work, we study one-shot video object localization problem that aims to localize instances of unseen objects in the target video using a single query image of the object. Toward addressing this challenging problem, we extend a popular and successful object detection method, namely DETR (Detection Transformer), and introduce a novel approach – query-guided detection transformer for videos (QDETRv). A distinctive feature of QDETRv is its capacity to exploit information from the query image and spatio-temporal context of the target video, which significantly aids in precisely pinpointing the desired object in the video. We incorporate cross-attention mechanisms that capture temporal relationships across adjacent frames to handle the dynamic context in videos effectively. Further, to ensure strong initialization for QDETRv, we also introduce a novel unsupervised pretraining technique tailored to videos. This involves training our model on synthetic object trajectories with an analogous objective as the query-guided localization task. During this pretraining phase, we incorporate recurrent object queries and loss functions that encourage accurate patch feature reconstruction. These additions enable better temporal understanding and robust representation learning. Our experiments show that the proposed model significantly outperforms the competitive baselines on two public benchmarks, VidOR and ImageNet-VidVRD, extended for one-shot open-set localization tasks.

Introduction

The field of computer vision has long been engaged in the pursuit of localizing objects of interest within videos. In the past, the primary emphasis has been localizing the specified object within a single frame on the entire video (Sivic and Zisserman 2003). Additionally, efforts have been made towards recognizing individuals based on their facial features (Sivic, Everingham, and Zisserman 2005) or clothing characteristics (Brunelli and Falavigna 1995). In a more novel direction, some research has explored using natural language or sketch-based queries to precisely locate objects within images (Minderer et al. 2022; Sadhu, Chen, and Nevatia 2019; Kumar and Mishra 2023; Tripathi et al. 2020, 2023). As visual comprehension continues to advance, there is a need for localization methodologies that can effectively pinpoint unfamiliar objects in the video that were not encountered during the training phase and unexpectedly appeared during testing. This is particularly relevant in various application areas, including surveillance systems and industrial automation. We illustrate our open-set one-shot object localization goal in Figure 1.

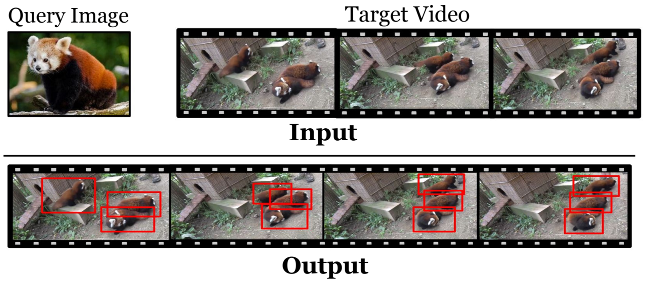

Figure 1: **Illustration of the proposed task.** Given an unseen query image and a target video, the model is expected to localize all instances of objects present in the query image on the target video. **[Best viewed in color]**.

To address the challenge posed by open-set one-shot object localization, our goal is to develop a model that, once trained, can localize any object query that might arise unexpectedly during testing. As noted in some of the recent works (Hsieh et al. 2019a; Osokin, Sumin, and Lomakin 2020), such a model should maintain its class-agnostic nature even during the training process. Our requisites, distinct from approaches, necessitate prior access to the list of category names during both training and testing (Fan, Tang, and Tai 2022; Du et al. 2022; Liang et al. 2023). To this end, we extend a popular and successful object detection method, namely DETR (Detection Transformer) (Carion et al. 2020), and introduce a novel query-guided detection transformer for videos – QDETRv. The QDETRv leverages the visual representation of the query image to better localize the target object in the video frames. To handle the temporal context in videos, we incorporate cross-attention that captures information from neighboring frames of the video. The key and value matrices of the cross-attention are obtained from the contextual and target frames, and the query matrix is derived from the query image. The cross-attention operation produces the final context-aware feature representation, which inputs the query-guided DETR. To ensure stronger initialization of QDETRv for our targeted task, we introduced an unsupervised, video-specific pretraining approach. During the pretraining phase, we employed a feature reconstruction loss combined with recurrent object queries to enhance representation quality and temporal learning.

To facilitate our study, we have extended two existing datasets, VidOR (Shang et al. 2019a) and ImageNet-VidVRD (Shang et al. 2017a), by splitting them into train and test sets such that there are no common object categories between the two sets. We obtain query images from each object category from the Google open-images (Kuznetsova et al. 2018) dataset. This dataset preparation ensures a realistic evaluation of one-shot learning performance on videos containing unseen objects. Here, we must highlight that we employ category-wise splits to conduct experiments. However, our model remains entirely class-agnostic during the training and testing phases. Our experimental results demonstrate the effectiveness of our proposed one-shot video object localization approach on the extended VidOR (Shang et al. 2019a) and ImageNet-VidVRD (Shang et al. 2017a) datasets. By combining the query-guided DETR with the video-specific context module, our method exhibits robust performance in localizing instances of unseen objects in target videos.

The contributions of this work are twofold: (i) We proposed a novel category-free and query-guided extension to DETR. Our proposed model – QDETRv incorporates a temporal module, capitalizing on temporal context for superior object localization within videos. (ii) We introduced a novel video-specific unsupervised pretraining for QDETRv. This pretraining objective is analogous to our downstream one-shot localization task, resulting in a notable improvement in the localization performance.

Related Work

One/Few-Shot Object Detection

Over the last few years, there has been significant progress in one/few-shot object detection, largely due to the adoption of sophisticated strategies such as attention mechanisms, transformers, and few-shot learning techniques (Sun et al. 2021; Fan et al. 2020; Wu et al. 2020; Kang et al. 2019; Wang et al. 2020; Sun et al. 2021; Fan, Tang, and Tai 2022). These strategies have been successful in a variety of domains, encompassing both image-based and video-based object detection, thus broadening the scope and application of these advanced detection methods. Recently, CoAE (Hsieh et al. 2019b) and OS2D (Osokin, Sumin, and Lomakin 2020) have shown significant advancements toward one-shot object detection in images. The CoAE (Hsieh et al. 2019b) makes ingenious use of co-attention and co-excitation mechanisms to generate region proposals and highlight correlated feature channels, thereby improving the accuracy and efficiency of object detection. On the other hand, the OS2D (Osokin, Sumin, and Lomakin 2020) model stands out as a versatile one-stage system that simultaneously performs localization and recognition tasks, proving its effectiveness in various detection scenarios. In the realm of few-shot object detection, recent innovations have resulted in the development of models like the feature reweighting-based model (Kang et al. 2019), the Multi-level Feature Enhancement (MFE) model (Wu et al. 2020), the Counting-DETR (Nguyen et al. 2022) model for few-shot object counting and detection, and the Fast Hierarchical Learning (She et al. 2022) model designed to address the catastrophic forgetting issue. These models, each exploring different aspects of few-shot learning, have contributed to improved performance across various datasets and settings. In the context of video object detection, models like the Tube Proposal Network with Temporal Matching Network (Nguyen et al. 2022) and the Thaw method (Yu et al. 2022) for few-shot learning in video object detection have been proposed. Both models harness the power of attention mechanisms, transformers, and few-shot learning techniques to achieve compelling results. There have been several successful advances towards developing DETR (Carion et al. 2020) variants in recent years for addressing different objectives; for example, in (Dong et al. 2022), authors proposed an incremental DETR that can generalize to novel classes with finetuning on a few examples. In (Jia et al. 2022) proposed different variants of object query, and (Jia et al. 2022) proposed an auxiliary one-to-many matching branch during training to enhance the performance of DETR for object detection in images. In (Dai et al. 2021) proposed UP-DETR method that enhances the original DETR (Carion et al. 2020) model by introducing unsupervised pretraining with a random query patch detection pretext task, demonstrating considerable performance improvement. The advancements in these areas suggest that the continuous refinement of these techniques can lead to substantial improvements in object detection across image and video domains with limited supervision.

Transformer-Based Pretraining

Pretraining methods based on transformers, such as BERT (Kenton and Toutanova 2019) and GPT (Brown et al. 2020) for natural language processing, and ViT (Dosovitskiy et al. 2021) for computer vision, have revolutionized the landscape of various tasks, including object detection. These methods leverage self-attention mechanisms to learn high-level semantic features from large-scale data, resulting in models with robust generalization capabilities. By pretraining on large-scale data with diverse modalities, such as image-text pairs (Minderer et al. 2022), these models can enhance the quality of representation learning, and also enable the transfer of this knowledge to downstream tasks, even when object-level data are scarce, thereby circumventing some of the limitations of traditional methods. Furthermore, unsupervised pretraining techniques, such as the random query patch detection proposed by UP-DETR (Dai et al. 2021), have shown promise in enhancing the performance of object detection models. By capitalizing on the strengths of the transformer architecture in capturing long-range dependencies and spatial localization, these techniques offer exciting avenues for future research and development in object detection.

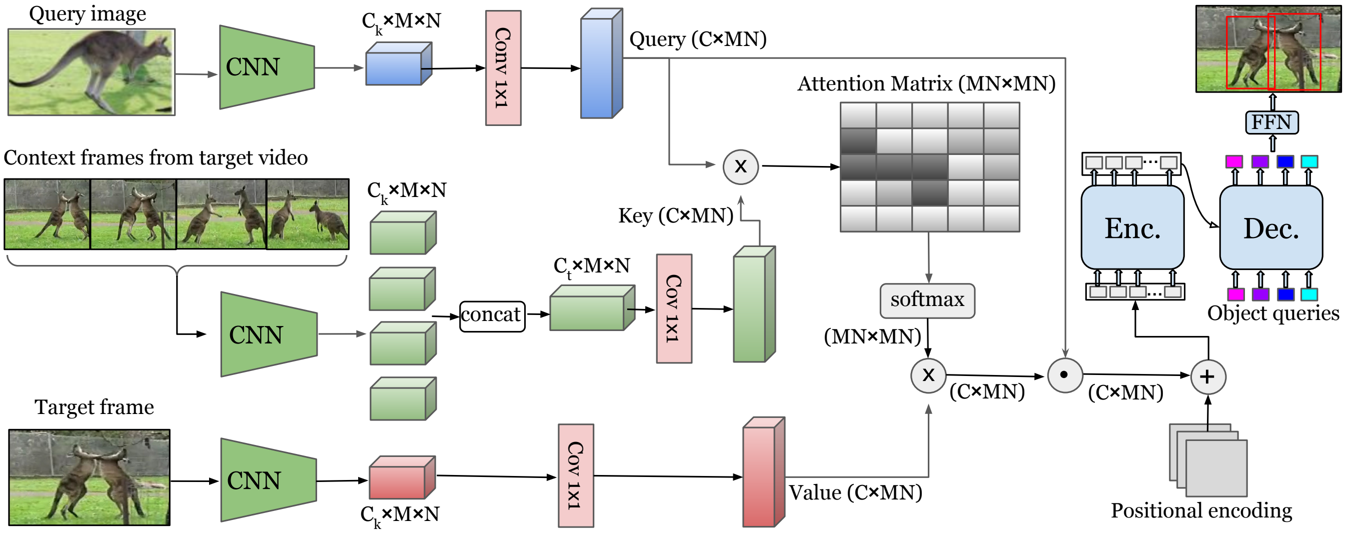

Figure 2: **Illustration of the proposed QDETRv.** The process begins with the feature extraction of a query image and video frames using a CNN encoder. A cross-attention mechanism and dot-product attention are used to create an attention map, transforming target frame features. The output is integrated into DETR's encoder, and predictions for bounding boxes are generated using the DETR decoder. **[Best viewed in color]**.

Query-Guided DETR for Videos (QDETRv)

Given a query image I_q featuring an unseen object, the objective is to spatiotemporally localize all instances of the unseen object within a target video V. Videos, with their intricate interplay of spatial and temporal dimensions, present both opportunities and challenges for localization. Harnessing this complexity effectively ensures high precision and reduces false localization. To this end, in this work, we proposed QDETRv that effectively harnesses the temporal context from the neighboring frames for better localization of unseen objects.

In this work, we utilize a query image I_q, a frame I_t, and a set of adjacent frames S from the target video. These frames are defined as $S = \{I_{t+j}\}$, where $j = -k \ to \ k$ with temporal context window $2k + 1$ for the t^{th} frame of the video V.

Feature Encoding: The query image, target, and contextual frames are processed through a Convolutional Neural Network (CNN) encoder to generate corresponding feature representations. This results in the dimensions $I_q, I_t \in \mathbb{R}^{C_k \times M \times N}$ and $S \in \mathbb{R}^{C_t \times M \times N}$, where M and N are the height and width of the CNN feature maps and C_k, and C_t are the channel dimensions.

Cross-Attention Mechanism for Temporal Context: In our work, we utilize the cross-attention mechanism (Vaswani et al. 2017) to encode temporal context within the target frame and enable interaction between the query image and the target frame. As part of this process, we first generate Query $Q \in \mathbb{R}^{C \times MN}$, Key $K \in \mathbb{R}^{C \times MN}$, and Value $V \in \mathbb{R}^{C \times MN}$ matrices using convolution operations of kernel size 1×1. Subsequently, we apply the dot-product attention (Vaswani et al. 2017) between Query and Key, which results in an attention map. The latter serves as a storage of correspondences between the visual feature of the query image and the temporal context extracted from the video. This attention map is then used to transform the target frame features represented by Value. This process leads to output features $O \in \mathbb{R}^{C \times MN}$, which can be defined by the following equations:

$$W = \text{Softmax}\left(\frac{Q^T K}{\sqrt{C}}\right), \tag{1}$$

$$O = VW^T \odot Q. \tag{2}$$

Here, $\odot$ represents element-wise multiplication. The output features are enriched by adding positional embedding and are then integrated into the encoder of DETR (Carion et al. 2020). This allows the localization of unseen objects within the target frame, which is treated as a direct set problem similar to DETR (Carion et al. 2020). The decoder's role is to generate N pairs of predictions $y = \{\hat{y_i}\}_{i=1}^N$ to accurately determine the bounding box of the query object in the current frame. Further, we leverage the Hungarian algorithm to calculate the matching cost between the prediction $\hat{y}_{\hat{\sigma}(i)}$ and the ground truth y_i. Here, $\hat{\sigma}(i)$ signifies the index of y_i as computed by optimal bipartite matching.

Loss Formulation: The predicted result $\hat{y_i} = (\hat{c_i} \in \mathbb{R}^2, \hat{b_i} \in \mathbb{R}^4, \hat{p_i} \in \mathbb{R}^C)$ comprises three elements: (a) $\hat{c_i}$, this binary classification determines whether a match with the query object is found ($c_i = 1$) or not ($c_i = 0$) for each object query. (b) $\hat{b_i}$, this vector is responsible for defining the box center coordinates along with its width and height, represented as x, y, w, h. (c) $\hat{p_i}$ is feature reconstruction used during pretraining for additional supervision. Taking into account these definitions, we define the Hungarian loss for all matched pairs as:

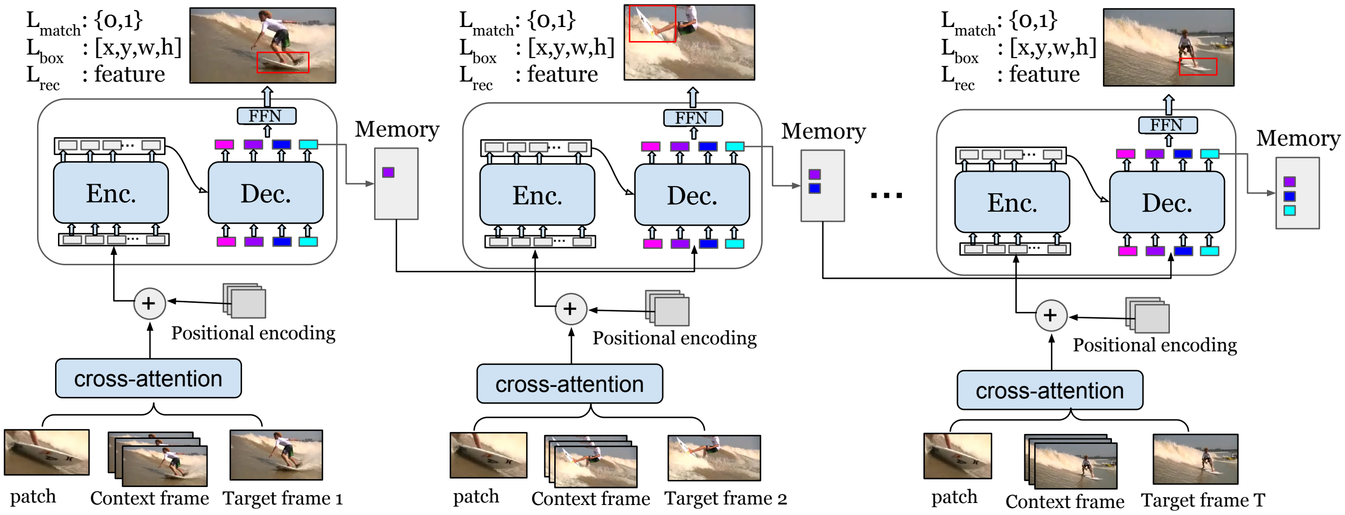

Figure 3: **Unsupervised Video Pretraining:** Each frame, along with its associated query patch and context frame, is fed into QDETRv. The model detects the patch on every frame using decoded object queries responsible for predicting the bounding box. These decoded queries are then forwarded to the next frame as object queries for the decoder. Containing the semantic information extracted from previous frames for patch objects, they facilitate the detection in subsequent frames. During this phase, an unsupervised query reconstruction loss is also applied as an extra supervisory mechanism. The model has been pre-trained on a synthetically created dataset to enhance its performance. **[Best viewed in color].**

$$\mathcal{L}(y,\hat{y}) = \sum_{t=1}^{T}\sum_{i=1}^{N}\Big[\lambda_{c_i}\mathcal{L}_{\text{match}}(c_i^t,\hat{c}^t_{\hat{\sigma}(i)}) + \mathbb{1}_{c_i=1}\mathcal{L}_{\text{box}}(b_i^t,\hat{b}^t_{\hat{\sigma}(i)}))\Big]. \tag{3}$$

In this equation, T is the number of frames, $\mathcal{L}_{match}$ refers to the binary cross-entropy loss over two classes (match the query object vs. not match), and λ symbolizes the class balance weight. The $\mathcal{L}_{box}$ is a combination of l_1 loss and the generalized IoU loss, with the weight hyper-parameters being the same as those in DETR (Carion et al. 2020).

Unsupervised Pretraining of QDETRv

Inspired by UP-DETR (Dai et al. 2021), we introduce a video-specific pretraining strategy suitable for QDETRv. This pretraining strategy is designed to locate a specific query patch within each frame, a process that aligns closely with our main objective. We devised a synthetic dataset based on the UCF101 (Soomro, Zamir, and Shah 2012) dataset inspired by (Wang et al. 2021). Our video-specific pretraining aims to identify and locate the query patch in every video frame, mimicking our main video frame localization task, as visualized in Figure 3. To ensure our pretraining approach is effective and aligns well with our primary task, we have added two additional mechanisms: **(i)** We integrated an unsupervised query reconstruction loss, as suggested in (Dai et al. 2021). This additional loss improves training signals and ensures high-quality embeddings are formed. The combined loss for this pretraining phase is represented as:

$$\mathcal{L}(y,\hat{y}) = \sum_{t=1}^{T}\sum_{i=1}^{N}\Big[\lambda_{c_i}\mathcal{L}_{\text{match}}(c_i^t,\hat{c}^t_{\hat{\sigma}(i)}) + \mathbb{1}_{c_i=1}\mathcal{L}_{\text{box}}(b_i^t,\hat{b}^t_{\hat{\sigma}(i)}) + \mathbb{1}_{c_i=1}\mathcal{L}_{rec}(p_i^t,\hat{p}^t_{\hat{\sigma}(i)})\Big]. \tag{4}$$

Here, $\mathcal{L}_{rec}$ represent the reconstruction loss defined below:

$$\mathcal{L}_{\text{rec}}(p_i^t,\hat{p}^t_{\hat{\sigma}(i)}) = \left\| \frac{p_i^t}{|p_i^t|_2} - \frac{\hat{p}^t_{\hat{\sigma}(i)}}{|\hat{p}_{\hat{\sigma}(i)}|_2} \right\|_2^2. \tag{5}$$

(ii) To enhance the temporal context further, we used recurrent object queries. We fed decoded object queries from the decoder (which localizes the query patch) into the decoder again for predicting the patch in the subsequent frame, as depicted in Figure 3. To incorporate recurrent queries, we used a memory module. The memory is a 256×100 matrix where 100 is the number of object queries used in DETR, with each object query having a dimension of 256×1. During pretraining, for every frame, it stores decoded object queries (recurrent queries) responsible for localization. The recurrent queries carry temporal context from the previous frames and help to localize the object in the subsequent frames. Their utility is quantitatively evaluated in Table 4.

We also incorporated an unsupervised image-level pretraining for UP-DETR (Dai et al. 2021). During this image-level pretraining phase, the primary objective is to localize a randomly selected patch within the input image. For this pretraining stage, we employed the ImageNet (Russakovsky et al. 2015) dataset.

Method	Pretraining	Vid-OR (Shang et al. 2019a)		VidVRD (Shang et al. 2017a)	
		Seen	Unseen	Seen	Unseen
CoAE (Hsieh et al. 2019a)	-	35.3	32.8	29.2	27.1
Retrieval-Based (Osokin, Sumin, and Lomakin 2020)	-	3.11	2.47	2.08	1.48
OWL-ViT (Minderer et al. 2022)	-	-	28.6	-	22.5
UP-DETR (Dai et al. 2021)	✗	37.4	35.4	33.7	28.5
UP-DETR (Dai et al. 2021)	✓	41.7	38.1	39.4	37.1
Ours					
QDETRv	✗	40.2	38.6	38.9	35.7
QDETRv	✓	**43.1**	**41.6**	**42.6**	**38.5**

Table 1: Comparative performance of different baselines and our method on Vid-OR (Shang et al. 2019a) and VidVRD (Shang et al. 2017a) datasets. The metric for evaluation is mAP. QDETRv methods outperform other baselines across both datasets and splits, with pertaining significantly boosting the performance.

Split	Vid-OR		VidVRD	
	Seen	Unseen	Seen	Unseen
#Train Videos	4,795	6,164	574	758
#Train Object Queries	53,277	57,599	3,413	4,395
#Train Object Categories	75	75	30	30
#Test Videos	1,319	32	184	42
#Test Object Queries	29,108	407	1,088	112
#Test Object Categories	75	9	30	5

Table 2: Overview of the Vid-OR (Shang et al. 2019a) and VidVRD (Shang et al. 2017a) datasets, summarizing the distribution of training and testing videos, object queries, and object categories for both seen and unseen splits.

Experiments and Results

Datasets and Performance Measure

In this work, we employed three primary datasets: VidOR (Shang et al. 2019b), ImageNet-VidVRD (Shang et al. 2017b), and Open Images (Kuznetsova et al. 2018). We further refined VidOR and ImageNet-VidCRD annotations, segregating super-classes like *fruits* and *vegetables*. Extracted target frames from both datasets were paired with Open Images. Some low-frequency classes were excluded during training, but all frames were used in testing. Note that our method is class-agnostic and does not use the class information during training and testing. Dataset statistics are provided in Table 2.

Dataset for Pretraining: For effective visual representation learning, synthetic datasets are crucial given the challenging task of annotating real object movements in videos. We begin by generating a simulated trajectory resembling object movement. Consistency is ensured using bounding boxes at chosen keyframes, with in-between positions filled by linear interpolation. Patches from real video frames are then superimposed on this trajectory. The UCF101 (Soomro, Zamir, and Shah 2012) dataset, featuring 13,320 videos with diverse complexities, was employed for pretraining, given its suitability to improve our approach.

Performance Measure: We have utilized the mean Average Precision (mAP) at an Intersection over Union (IoU) of 0.5 as our primary evaluation metric. Our evaluation operates framewise, considering each frame as a separate detection problem.

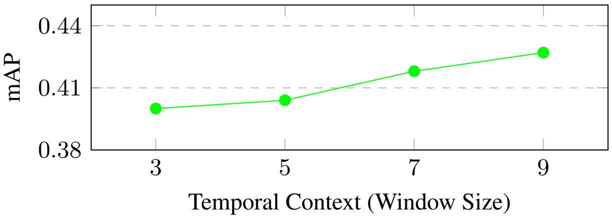

Figure 4: Ablation study on the temporal context window size using the VidOR dataset. A positive correlation is observed between the window size and the mAP.

Baselines and Implementation Details

The absence of prior work on class-agnostic one-shot video-based object localization methods inclines us to benchmark our proposed method against image-based techniques in a frame-wise fashion. However, such models might not be able to capture temporal aspects of the video.

Retrieval-based approach: A naive approach would involve combining an object detector that detects all objects as one class coupled with an image retrieval system for one-shot detection. This system utilizes the object detector's detections as a database and class objects as queries to search for relevant images. Inspired by baselines in (Osokin, Sumin, and Lomakin 2020), we utilized object detectors with identical architectures.

CoAE (Hsieh et al. 2019a): It uses non-local operations to explore the co-attention embodied in each query-target pair and yield region proposals accounting for the one-shot situation. It then formulates a squeeze-and-co-excitation that adaptively emphasizes correlated feature channels to help uncover relevant proposals and, eventually, the target objects. This method intentionally casts the learning formulation such that it does not solely rely on the label information of training data but instead explores correlated evidence revealed by the query-target pairs.

Method	Bear	Carrot	Coconut	Frisbee	Kangaroo	Lemon	Leopard	Melon	Orange	Stop sign
CoAE	37.8	19.2	33.6	25.4	42.6	21.3	46.9	22.8	28.1	30.6
OWL-ViT	35.7	20.3	28.5	21.7	38.6	22.8	42.8	20.3	25.3	27.4
UP-DETR	47.5	32.6	36.1	32.8	50.1	30.2	50.7	31.4	30.2	35.3
QDETRv	**47.9**	**35.3**	**38.6**	**33.4**	**51.4**	**30.9**	**52.3**	**33.2**	**30.8**	**35.8**

Table 3: Class-wise performance comparison. Each column indicates performance for distinct object categories. Larger objects like bears, kangaroos, and leopards are better detected than smaller ones like coconuts and lemons.

Recurrent Query	VidVRD	VidOR
✗	37.9	40.3
✓	**38.5**	**41.6**

Table 4: Ablation studies on recurrent queries during pretraining showed improved performance. This highlights the importance of adding temporal context in the pretraining phase.

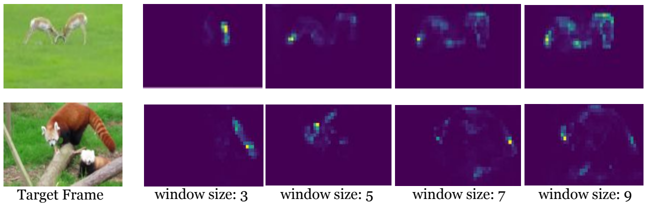

Figure 5: Cross-attention visualization on test frame with varying window size. Attention improves on objects of interest with an increase in window size.

OWL-ViT (Minderer et al. 2022): We used it as the one-shot image-conditioned object detection method for the video by applying it at each frame. OWL-ViT used a Vision Transformer architecture and contrastive image-text pretraining.

UP-DETR (Dai et al. 2021): It refines the DETR model via unsupervised pretraining using a random query patch detection task. By selecting random image patches as queries and training the model to detect them, UP-DETR addresses challenges like balancing classification with localization by freezing the CNN backbone and managing multiple query patches with attention masks. This approach significantly boosts UP-DETR's performance in one-shot object detection, leveraging pretraining to achieve enhanced accuracy with minimal labeled data.

Implementation Details: We pre-train and fine-tune our models using the Adam optimizer (Kingma and Ba 2015) with an initial learning rate = 1e-5. The initialization of our frame and query encoders leverages weights of ResNet-50 pre-trained on the ImageNet dataset (Russakovsky et al. 2015). We train the model for 200 epochs with batch size = 350. Our implementation was done using the PyTorch library. We trained the model on three Nvidia-RTX A6000 GPUs.

Results and Discussions

The results presented in Table 1 showcase the performance comparison of various methods applied for both of the datasets, Vid-OR (Shang et al. 2019a) and Vid-VRD (Shang et al. 2017a). The methods assessed include baselines Retrieval-Based (Osokin, Sumin, and Lomakin 2020), OWL-ViT (Minderer et al. 2022), CoAE (Hsieh et al. 2019b), UP-DETR (Dai et al. 2021) (QDETRv without temporal module), and our QDETRv, with the latter two being tested both with and without pretraining.

Performance of QDETRv: For the Vid-OR dataset, QDETRv with video pretraining yielded the best performance with 43.1% for seen and 41.6% for unseen classes, surpassing its performance without pretraining, which scored 38.6% and 40.2%, respectively. It's important to note that unsupervised video pretraining improved the results significantly, highlighting its effectiveness as a strategy for this task. We get similar observations for the ImageNet-VidVRD dataset.

Ablation Study on Temporal Context Module: To evaluate the effectiveness of the temporal context module integrated into QDETRv, we conducted an ablation study where this module was removed. We exclusively adopted an image-level approach for unsupervised pretraining, eliminating the recurrent queries and leveraging the ImageNet (Russakovsky et al. 2015) dataset. Without a context module with image-level pretraining, our method becomes similar to UP-DETR (Dai et al. 2021). The results from this ablation demonstrated a notable 3.2% performance boost for QDETRv without pretraining. Furthermore, with pretraining, there was a 1.4% enhancement in performance on VidOR (Shang et al. 2019a) dataset for the unseen test set as shown in Table 1.

Fine-tuning with varied Temporal Context Windows: As we progressed to the fine-tuning stage, we explored different temporal context windows, spanning a range from 3 to 9, as depicted in Figure 4. Interestingly, our results indicated that as the temporal context window expanded, there was a corresponding enhancement in performance. This trend underscores the pivotal role of temporal context in improving localization accuracy.

Cross-attention Visualization for varied temporal context: Figure 5 shows the cross-attention visualization on the test frame for different window sizes. The attention to the object of interest improves with an increase in the window size, showing the efficacy of temporal context being utilized from context frames of the video.

Ablation on Recurrent Queries: Table 4 presents an abla-

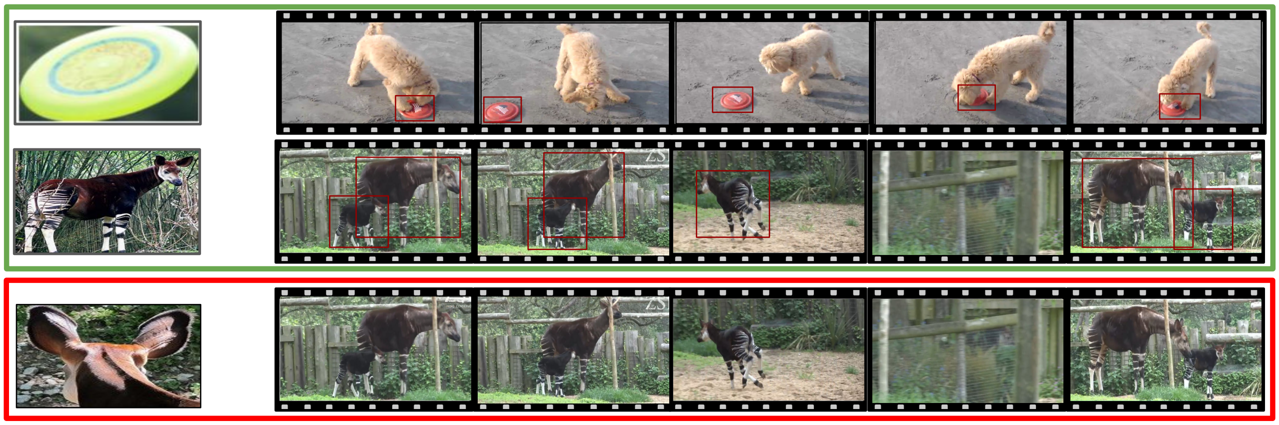

Figure 6: **Qualitative results.** In the green box results, QDETRv accurately localizes objects from the query image on the left. The red box highlights the model's limitations, with missed localization in videos where the object is only partially visible in the query image.

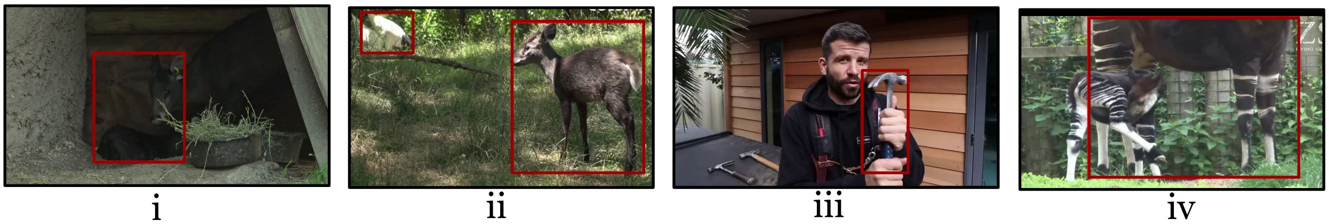

Figure 7: **Failure Cases.** We observed the following major failure types: (i) The occluded object is missed, (ii) A visually similar object is detected as dear on the top left, (iii) a few small instances of the hammer are missed, (iv) an overlapping instance is commonly detected as a single object.

tion study focusing on the impact of recurrent queries during the pretraining phase. When recurrent queries are employed (indicated by ✓), there's a noticeable performance improvement on both datasets: 0.6% on ImageNet-VidVRD (Shang et al. 2017a) and 1.3% on VidOR (Shang et al. 2019a), compared to when these queries are omitted (✗). While seemingly modest, these increments emphasize the added value of recurrent queries in capturing intricate temporal information during pretraining.

Class-wise Performance: Table 3 shows the class-wise performance on VidOR dataset. Intriguingly, larger-sized objects like bear, kangaroo, and leopard generally achieve better performance compared to relatively smaller objects, such as coconut and lemon.

Qualitative Results: QDETRv effectively localizes query objects both spatially and temporally within the videos, as shown in the first two rows of Figure 6. Notably, in the second video, QDETRv adeptly identifies multiple instances of the object.

Failure Case Analysis: While our proposed methodology has shown promising results, it does come with a few limitations: our model QDETRv highly relies on the quality of the query image for object localization in videos. As shown in Figure 6 (red box), occluded query images often lead to weak or missing localization. Further, we observe the following major failure cases: (i) occlusion, (ii) visually similar object, (iii) small object, and (iv) overlapping instances, on randomly selected video frames. Figure 7 shows a selection of these examples. We leave addressing these limitations of the current model as the future scope of this work.

Conclusions

We introduced the novel class-agnostic query-guided DETR for videos (QDETRv), effectively enhancing video object localization using a query image and temporal context. We showed evaluations on two extended datasets that displayed superior and consistent performance of QDETRv, particularly when paired with video-specific unsupervised pretraining. While showing promise, our approach has a few limitations, including limited success in localizing small or occluded object instances. Further, it struggles to differentiate between visually similar objects. We leave addressing these limitations as future work and firmly believe this work will open up new research avenues in open-set one-shot video object localization.

Acknowledgment: This work is supported by a gift grant from Accenture Labs (project number: S/AC-T/AM/20220078). Yogesh Kumar is supported by a UGC fellowship.

References

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. In *Proc. NuerIPS*.

Brunelli, R.; and Falavigna, D. 1995. Person identification using multiple cues. *PAMI*, 17.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *Proc. ECCV*.

Dai, Z.; Cai, B.; Lin, Y.; and Chen, J. 2021. UP-DETR: Unsupervised pre-training for object detection with transformers. In *Proc. CVPR*.

Dong, N.; Zhang, Y.; Ding, M.; and Lee, G. H. 2022. Incremental-DETR: Incremental Few-Shot Object Detection via Self-Supervised Learning. In *Proc. AAAI*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *Proc. ICLR*.

Du, X.; Wang, X.; Gozum, G.; and Li, Y. 2022. Unknown-Aware Object Detection: Learning What You Don't Know from Videos in the Wild. In *Proc. CVPR*.

Fan, Q.; Tang, C.-K.; and Tai, Y.-W. 2022. Few-shot video object detection. In *Proc. ECCV*.

Fan, Q.; Zhuo, W.; Tang, C.-K.; and Tai, Y.-W. 2020. Few-shot object detection with attention-RPN and multi-relation detector. In *Proc. CVPR*.

Hsieh, T.-I.; Lo, Y.-C.; Chen, H.-T.; and Liu, T.-L. 2019a. One-Shot Object Detection with Co-Attention and Co-Excitation. In *Proc. NeurIPS*.

Hsieh, T.-I.; Lo, Y.-C.; Chen, H.-T.; and Liu, T.-L. 2019b. One-shot object detection with co-attention and co-excitation. In *Proc. NuerIPS*.

Jia, D.; Yuan, Y.; He, H.; pei Wu, X.; Yu, H.; Lin, W.; huan Sun, L.; Zhang, C.; and Hu, H. 2022. DETRs with Hybrid Matching. *Proc. CVPR*.

Kang, B.; Liu, Z.; Wang, X.; Yu, F.; Feng, J.; and Darrell, T. 2019. Few-shot object detection via feature reweighting. In *Proc. CVPR*.

Kenton, J. D. M.-W. C.; and Toutanova, L. K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proc. NAACL-HLT*.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *Proc. ICLR*.

Kumar, Y.; and Mishra, A. 2023. Few-Shot Referring Relationships in Videos. In *Proc. CVPR*.

Kuznetsova, A.; Rom, H.; Alldrin, N. G.; Uijlings, J. R. R.; Krasin, I.; Pont-Tuset, J.; Kamali, S.; Popov, S.; Malloci, M.; Kolesnikov, A.; Duerig, T.; and Ferrari, V. 2018. The Open Images Dataset V4. *IJCV*, 128.

Liang, W.; Xue, F.; Liu, Y.; Zhong, G.; and Ming, A. 2023. Unknown Sniffer for Object Detection: Don't Turn a Blind Eye to Unknown Objects. In *Proc. CVPR*.

Minderer, M.; Gritsenko, A.; Stone, A.; Neumann, M.; Weissenborn, D.; Dosovitskiy, A.; Mahendran, A.; Arnab, A.; Dehghani, M.; Shen, Z.; Wang, X.; Zhai, X.; Kipf, T.; and Houlsby, N. 2022. Simple Open-Vocabulary Object Detection. In *Proc. ECCV*.

Nguyen, T.; Pham, C.; Nguyen, K.; and Hoai, M. 2022. Few-Shot Object Counting and Detection. In *Proc. ECCV*.

Osokin, A.; Sumin, D.; and Lomakin, V. 2020. Os2d: One-stage one-shot object detection by matching anchor features. In *Proc. ECCV*.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; Berg, A. C.; and Fei-Fei, L. 2015. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 211–252.

Sadhu, A.; Chen, K.; and Nevatia, R. 2019. Zero-shot grounding of objects from natural language queries. In *Proc. CVPR*.

Shang, X.; Di, D.; Xiao, J.; Cao, Y.; Yang, X.; and Chua, T.-S. 2019a. Annotating Objects and Relations in User-Generated Videos. In *Proc. ICMR*.

Shang, X.; Di, D.; Xiao, J.; Cao, Y.; Yang, X.; and Chua, T.-S. 2019b. Annotating Objects and Relations in User-Generated Videos. In *Proc. ICMR*.

Shang, X.; Ren, T.; Guo, J.; Zhang, H.; and Chua, T.-S. 2017a. Video Visual Relation Detection. In *ACM International Conference on Multimedia*.

Shang, X.; Ren, T.; Guo, J.; Zhang, H.; and Chua, T.-S. 2017b. Video Visual Relation Detection. In *ACM International Conference on Multimedia*.

She, Y.; Bhat, G.; Danelljan, M.; and Yu, F. 2022. Fast Hierarchical Learning for Few-Shot Object Detection. In *Proc. IROS*.

Sivic; and Zisserman. 2003. Video Google: A text retrieval approach to object matching in videos. In *Proc. ICCV*.

Sivic, J.; Everingham, M.; and Zisserman, A. 2005. Person spotting: video shot retrieval for face sets. In *Proc. CIVR*.

Soomro, K.; Zamir, A. R.; and Shah, M. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

Sun, B.; Li, B.; Cai, S.; Yuan, Y.; and Zhang, C. 2021. Fsce: Few-shot object detection via contrastive proposal encoding. In *Proc. CVPR*.

Tripathi, A.; Dani, R. R.; Mishra, A.; and Chakraborty, A. 2020. Sketch-guided object localization in natural images. In *Proc. ECCV*.

Tripathi, A.; Dani, R. R.; Mishra, A.; and Chakraborty, A. 2023. Multimodal query-guided object localization. *Multimedia Tools and Applications*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Proc. NeurIPS*.

Wang, G.; Zhou, Y.; Luo, C.; Xie, W.; Zeng, W.; and Xiong, Z. 2021. Unsupervised visual representation learning by tracking patches in video. In *Proc. CVPR*.

Wang, X.; Huang, T.; Gonzalez, J.; Darrell, T.; and Yu, F. 2020. Frustratingly Simple Few-Shot Object Detection. In *Proc. ICML*.

Wu, J.; Liu, S.; Huang, D.; and Wang, Y. 2020. Multi-scale positive sample refinement for few-shot object detection. In *Proc. ECCV*.

Yu, Z.; Wang, G.; Chen, L.; Raschka, S.; and Luo, J. 2022. When Few-Shot Learning Meets Video Object Detection. In *Proc. ICPR*.